UNIVERSIDAD CARLOS III DE MADRID

working papers

Working Paper 13-12
Statistics and Econometrics Series 11
June 2013

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-48

Bayesian multivariate Bernstein polynomial density estimation

Yanyun Zhao *, María Concepción Ausín † and Michael Peter Wiper ‡

Abstract

This paper introduces a new approach to Bayesian nonparametric inference for densities on the hypercube, based on the use of a multivariate Bernstein polynomial prior. Posterior convergence rates under the proposed prior are obtained. Furthermore, a novel sampling scheme, based on the use of slice sampling techniques, is proposed for estimation of the posterior predictive density. The approach is illustrated with both simulated and real data examples.

Keywords: Bayesian nonparametrics, Bernstein polynomials, Dirichlet process.

*Departamento de Estadística, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), Spain; Email: yanyun.zhao@uc3m.es.

†Departamento de Estadística, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), Spain; Email: concepcion.ausin@uc3m.es.

‡Departamento de Estadística, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), Spain; Email: michael.wiper@uc3m.es.

Bayesian multivariate Bernstein polynomial density estimation

Yanyun Zhao, María Concepción Ausín and Michael Peter Wiper
Departamento de Estadística, Universidad Carlos III de Madrid

Abstract

This paper introduces a new approach to Bayesian nonparametric inference for densities on the hypercube, based on the use of a multivariate Bernstein polynomial prior. Posterior convergence rates under the proposed prior are obtained. Furthermore, a novel sampling scheme, based on the use of slice sampling techniques, is proposed for estimation of the posterior predictive density. The approach is illustrated with both simulated and real data examples.

Keywords: Bayesian nonparametrics, Bernstein polynomials, Dirichlet process.

1 Introduction

Many real data samples possess characteristics such as multimodality, high skewness and kurtosis which are not well modeled by standard parametric distributions. In such cases, nonparametric modeling techniques might be preferable.

Although kernel density estimation techniques are the most popular approaches from the classical viewpoint, see e.g. Silverman (1986), in certain situations, alternative approaches based on approximating polynomials have been considered. In particular, Vitale (1975) developed a Bernstein polynomial based density estimator for density functions on a closed interval and this was extended to bivariate densities in Tenbusch (1994).

In the Bayesian context, most nonparametric density estimation is based on the use of Dirichlet process or Dirichlet process mixture priors, see e.g. Hjort et al. (2010) for a general review of the area. However, in the case of univariate densities on a closed interval, Petrone (1999a,b) develop an alternative approach based on the use of a Bernstein polynomial based prior. Consistency properties of the derived posterior distribution are examined in Petrone and Wasserman (2002) and the convergence rate of the posterior is derived in Ghosal (2001). An extended Bernstein polynomial prior model is examined in Trippa et al. (2011). Finally, software for Bayesian Bernstein polynomial density estimation was developed in Jara et al. (2011).

To the best of our knowledge however, Bernstein polynomial priors have not been generalized to the case of multivariate density estimation on the hypercube. Therefore, the main objectives of the present paper is to define the multivariate Bernstein polynomial prior distribution and to derive the convergence rate of posterior distribution of a multivariate Bernstein polynomial model under very general conditions. Furthermore, we also introduce a computational approach to implementing Bernstein polynomial density estimation which is based on the slice sampling algorithm for sampling Dirichlet process mixture models developed in Walker (2007) which is somewhat faster than the algorithm used in Petrone (1999a).

The rest of this paper is organized as follows. Firstly, in Section 2, we briefly outline the properties of both univariate and multivariate Bernstein polynomials. Secondly, in Section 3 we introduce a multivariate Bernstein polynomial prior and derive the associated posterior convergence rates. In Section 4, we provide an appropriate algorithm for sampling from the posterior parameter distribution. Section 5 then illustrates our approach with both simulated and real data examples and finally, some conclusions and extensions are provided in Section 6.

2 Bernstein polynomials

In this section we introduce Bernstein polynomials, which are well known to provide good approximations to continuous functions on a closed interval. Then, we illustrate the use of Bernstein polynomials to approximate distribution and density functions for variables defined on such an interval. More details and further results on the approximation properties of Bernstein polynomials are provided in e.g. Lorentz (1986) and Phillips (2003).

2.1 Univariate Bernstein polynomials

Let $g(x)$ be a continuous and bounded, real function defined on $[0,1]$. Then the Bernstein polynomial of degree k for $g(x)$ is defined by:

$$B(x;k,g) = \sum_{j=0}^{k} g\left(\frac{j}{k}\right) \binom{k}{j} x^j (1-x)^{k-j}. \tag{1}$$

Then, it is well known that, letting k tend to infinity, the Bernstein polynomial approximations converge uniformly to g and, moreover, that their derivatives also converge to the corresponding derivatives of g.

Then, in particular if F be a distribution function on $[0,1]$, then, it is easy to show that a corresponding, k'th order, Bernstein polynomial approximation to the corresponding density function is given by:

$$\begin{array}{rcl} b(x|k,F) & \stackrel{def}{=} & \frac{\partial}{\partial k} B(x;k,F) \\ & = & \sum_{j=1}^{k} \left(F\left(\frac{j}{k}\right) - F\left(\frac{j-1}{k}\right)\right) \beta(x|j,k-j+1), \end{array} \tag{2}$$

where $\beta(x|c,d) = \frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)} x^{c-1}(1-x)^{d-1}$ is a beta density.

2.2 Multivariate Bernstein polynomials

Let the m-dimensional unit hypercube be denoted by $[0,1]^m$. Then the associated m-dimensional Bernstein polynomial approximation at $\mathbf{x} = (x_1, \ldots, x_m)^T$, for a continuous, bounded function g on $[0,1]^m$ is defined by:

$$B(\mathbf{x}; k, g) = \sum_{j_1=0}^{k} \cdots \sum_{j_m=0}^{k} g\left(\frac{j_1}{k}, \ldots, \frac{j_m}{k}\right) \left(\prod_{r=1}^{m} \binom{k}{j_r} x_r^{j_r} (1-x_r)^{k-j_r}\right), \tag{3}$$

where $\mathbf{x} = (x_1, \ldots, x_m)^T$.

As in the univariate case, the Bernstein polynomials and their derivatives converge uniformly to g and their corresponding derivatives as $k \to \infty$. In the case that $g = F$ is a distribution function, then analogous to (2), the corresponding density approximation is given by:

$$\begin{array}{rcl} b(\mathbf{x}; k, F) & \stackrel{def}{=} & \dfrac{\partial^m B(\mathbf{x}; k, F)}{\partial x_1 \partial x_2 \ldots \partial x_m} \\ & = & \sum_{j_1=1}^{k} \cdots \sum_{j_m=1}^{k} w_{j_1 j_2 \ldots j_m} \prod_{r=1}^{m} \beta(x_r | j_r, k - j_r + 1) \end{array} \tag{4}$$

where $w_{j_1 j_2 \ldots j_m, k} = \int_{(j_1-1)/k}^{j_1/k} \cdots \int_{(j_m-1)/k}^{j_m/k} f(x_1, x_2, \ldots, x_m)\, dx_1 dx_2 \ldots dx_m$.

In particular, if the probability density $f(\cdot)$ is continuously differentiable on $[0,1]^m$ with bounded partial second derivative, then

$$\sup_{0 < x_1, x_2, \ldots, x_m \leq 1} |f(\mathbf{x}) - b(\mathbf{x}; k, F)| = O(k^{-1}). \tag{5}$$

This property can easily be shown by observing that

$$b(\mathbf{x}; k, F) = k^m \, \mathbb{E}\left(\int_{J_1/k}^{(J_1+1)/k} \cdots \int_{J_m/k}^{(J_m+1)/k} f(z_1, z_2, \ldots, z_m)\, \mathrm{d}z_1 \ldots \mathrm{d}z_m\right) \tag{6}$$

where $J_l \sim \text{Binomial}(k-1, x_l), l = 1, 2, \ldots, m$, when the result holds by Taylor series expansion.

3 Multivariate Bernstein prior distributions

In this section, we first define multivariate Bernstein polynomial prior distributions and then examine the convergence properties of the associated posterior distributions. The basic definition is a direct generalization of Petrone (1999b).

Let Δ be the space of distribution functions on $[0,1]^m$ and equip this with the Borel σ-field $\mathscr{F}$ generated by the topology of weak convergence. We need to construct a prior probability measure on $(\Delta, \mathscr{F})$ induced by the multivariate Bernstein polynomials.

A random distribution function H is a measurable map from a probability space $(\Omega, \mathscr{G}, \mathbf{P})$ to the space Δ. The distribution function of $\mathbf{P}H^{-1}$ is a prior probability measure on $(\Delta, \mathscr{F})$. Now we construct the triple $(\Omega, \mathscr{G}, \mathbf{P})$ as follows. Let $\mathbb{N}$ be the set of positive integers with power set $\mathscr{P}(\mathbb{N})$. Set $\Omega = (\mathbb{N} \times \Delta)$, $\mathcal{B}(\Omega)$ be the product σ-field $\mathscr{P}(\mathbb{N}) \times \mathscr{F}$, $\mathbf{P}$ the product measure and $\mathbf{x} \equiv (x_1, x_2, \ldots, x_m)$. Then define an operator from Ω to Δ by

$$B^*(\mathbf{x}; k, F) \equiv \begin{cases} 0 & \text{if } x_1 < 0 \text{ or } , \ldots , \text{ or } x_m < 0 \\ B(\mathbf{x}; k, F) & (x_1, x_2, \ldots, x_m) \in [0,1]^m \\ 1 & \text{in other cases} \end{cases}$$

Given (k, F), $B^*(\cdot; k, F)$ is a probability distribution function in $[0,1]^m$. $B^*(\cdot; k, F)$ is a random Bernstein polynomial if for each $(\mathbf{x}; k)$, $B(\mathbf{x}; k, \cdot)$ is a random variable from $(\Delta, \mathscr{F})$ in $\mathbb{R}$. So $B^*(\cdot; k, F)$ induces a probability measure π on $(\Delta, \mathscr{F})$. The full prior is then completed by setting prior distributions for k, F. Following Petrone (1999b), we shall assume that F follows a Dirichlet process prior, $F \sim \mathscr{D}(MF_0)$ and that $k \sim p(k)$ some prior probability distribution on $\mathbb{N}$ which we shall call a *Multivariate Bernstein Dirichlet prior*. Following Petrone (1999b), it is easy to show that this prior structure has full support and weakly converges to F as $k \to \infty$. In the following, we derive the convergence rate of the posterior distribution.

3.1 The convergence rate of the posterior distribution

For a distance d on a class of densities $\mathscr{F}$, let $D(\varepsilon, \mathscr{F}, d)$ stand for the ε-packing number defined to be the maximum number of points in $\mathscr{F}$ such that the distance between each pair is at least ε. Let the true density $f_0 \in \mathscr{F}$, a class of densities and let P_0 be the probability measure with density f_0. Let $\|f - f_0\|_1$ stand for the L_1-distance and $h(f, f_0) = \|f^{1/2} - f_0^{1/2}\|_2$ stand for the Hellinger distance. Define:

$$\begin{aligned} K(f_0, f) &= \int \log(f_0/f) \mathrm{d}P_0 \\ V(f_0, f) &= \int (\log(f_0/f))^2 \mathrm{d}P_0 \\ N(\varepsilon, f_0) &= \{f : K(f_0, f) \leq \varepsilon^2, V(f_0, f) \leq \varepsilon^2\} \end{aligned}$$

Let d stand for either the L_1-distance or the Hellinger distance. Then the following theorem from Ghosal (2001) is an important starting point for our paper.

Theorem 3.1 *[Ghosal 2001] Let Π_n be a sequence of priors on $\mathscr{F}$. Suppose that for positive sequences $\bar{\varepsilon}_n$, $\tilde{\varepsilon}_n \to 0$ with $n \min(\bar{\varepsilon}_n^2, \tilde{\varepsilon}_n^2) \to \infty$, constants $c_1, c_2, c_3, c_4 > 0$ and sets $\mathscr{F}_n \subset \mathscr{F}$, we have*

$$\log D(\tilde{\varepsilon}_n, \mathscr{F}_n, d) \leq c_1 n \bar{\varepsilon}_n^2, \tag{7}$$

$$\Pi_n(\mathscr{F} \backslash \mathscr{F}_n) \leq c_3 \exp(-(c_2 + 4) n \tilde{\varepsilon}_n^2), \tag{8}$$

$$\Pi_n(N(\tilde{\varepsilon}_n, f_0) \geq c_4 \exp(-c_2 n \tilde{\varepsilon}_n^2). \tag{9}$$

Then for $\varepsilon_n = \max(\bar{\varepsilon}_n, \tilde{\varepsilon}_n)$ and a sufficiently large $M > 0$, the posterior probability

$$\Pi_n(f : d(f, f_0) > M \varepsilon_n | X_1, \ldots, X_n) \to 0 \text{ in } P_0^n\text{-probability.}$$

Denote $\tilde{k} \equiv k^m$. Let $Q(\tilde{k}; \alpha_{1,\tilde{k}}, \ldots, \alpha_{\tilde{k},\tilde{k}})$ be the probability measure induced on $\mathscr{B}_{\tilde{k}}$ by assigning the Dirichlet distribution $D(\tilde{k}; \alpha_{1,\tilde{k}}, \ldots, \alpha_{\tilde{k},\tilde{k}})$ to $\mathbf{W}_{\tilde{k}}$, where $\mathscr{B}_{\tilde{k}}$ is the class of all Bernstein densities of order k on $[0,1]^m$. The prior on the density f is then the mixture

$\sum_{i_1=1}^{k} \cdots \sum_{i_m=1}^{k} \rho(\tilde{k}) Q(\tilde{k}; \alpha_{1,\tilde{k}}, \ldots, \alpha_{\tilde{k},\tilde{k}})$.

The next result shows that the rate of convergence $n^{-1/2} \log n$ is obtained when the true density is actually a Bernstein density. We omit the proof since it is almost identical to that of Theorem 2.4 of Ghosal (2001).

Theorem 3.2 *[Ghosal 2001] Let the true density $f_0 = b(\cdot, \cdots, \cdot; k_0, \mathbf{w}_{k_0}^0)$ for some k_0 and $\mathbf{w}_{k_0} \in \Delta_{k_0^m}$. Let $0 < \rho(k) \leq B \exp(-\beta k)$ for some constants B and β. Then for a sufficiently large constant C,*

$$\Pi(f : d(f, f_0) > C \frac{\log n}{\sqrt{n}} | X_1, \ldots, X_n) \to 0 \text{ in } P_0^n\text{-probability.}$$

When the true density f_0 is not of the Bernstein type, the convergence rate will naturally be much slower. A proof of the following result, which generalizes Theorem 3.1 of Ghosal (2001) to the multivariate case and shows that the posterior distribution converges at the rate $n^{-1/(m+2)} (\log n)^{(m+4)/(2m+4)}$, is given in the Appendix.

Theorem 3.3 *Let the true density f_0 be bounded away from 0 and have bounded second derivative. Consider a Bernstein polynomial prior for f satisfying the condition $B_1 e^{-\beta_1 j} \leq \rho(j) \leq B_2 e^{-\beta_2 j}$ for all j for some constants $B_1, B_2, \beta_1, \beta_2 > 0$. then for a sufficiently large constant C,*

$$\Pi(f : d(f, f_0) > C n^{-1/(m+2)} (\log n)^{(m+4)/(2m+4)} | X_1, \ldots, X_n) \to 0 \text{ in } P_0^n\text{-probability.}$$

4 Computational implementation

Suppose now that we have a data sample $\mathbf{X} = \mathbf{x_1}, \ldots, \mathbf{x_n}$, where $\mathbf{x_i} = (x_{i1}, \ldots, x_{im})^T$. Given a multivariate Bernstein Dirichlet prior, in particular, we are interested in sampling from the predictive density of new observation. Note first that we can represent the model structure in a hierarchical way as follows:

$$
\begin{aligned}
k &\sim p(k) \\
F|k &\sim DP(M, F_0) \\
\mathbf{X}_i|k, F &\sim b(\cdot; k, F) \quad \text{are conditionally i.i.d.}
\end{aligned} \tag{10}
$$

where $b(\mathbf{x}; k, F)$ is as in (4) and the set of weights, $\mathbf{W}_k = (w_{11\ldots1,k}, w_{11\ldots2,k}, \ldots, w_{kk\ldots k,k})^T$ in $b(\cdot; k, F)$ has a Dirichlet distribution $D(.; \alpha_{11\ldots1,k}, \alpha_{11\ldots2,k}, \ldots, \alpha_{kk\ldots k,k})$, with parameters given by:

$$
\alpha_{j_1 j_2 \ldots j_m, k} = M \int_{(j_1-1)/k}^{j_1/k} \ldots \int_{(j_m-1)/k}^{j_m/k} f(\mathbf{x})\, dx_1 dx_2 \ldots dx_m,
$$

where $j_i = 1, \ldots, k \quad \text{and} \quad i = 1, \ldots, m.$

From Bayes theorem, the posterior distribution of $(k, \mathbf{W}_k|\mathbf{X})$ is proportional to:

$$
p(k) \frac{\Gamma(\alpha)}{\Gamma(\alpha_{11\ldots1,k}) \cdots \Gamma(\alpha_{kk\ldots k,k})} w_{11\ldots1,k}^{\alpha_{11\ldots1,k}} \cdots w_{kk\ldots k,k}^{\alpha_{kk\ldots k,k}} \prod_{i=1}^{n} b(x_{i1}, \ldots, x_{im}; k, F) \tag{11}
$$

Unfortunately, the computation of expression (11) is intractable. In the univariate case, Petrone (1999a,b) proposes a hybrid Monte Carlo algorithm to sample the posterior density. However, the computational time required by this algorithm is too high to be practicable in the multivariate case. Instead, we propose a slice sampling algorithm based on the use of auxiliary variables as the basis of a sampling scheme, as in Walker (2007), Papaspiliopoulos and Roberts (2008) .

From (10), we can use a stick breaking representation of the Dirichlet process to write the multivariate Bernstein Dirichlet prior as an infinite mixture as follows:

$$
f(\mathbf{x}_i|k, \boldsymbol{\rho}, \mathbf{y}) = \sum_{s=1}^{\infty} \rho_s \prod_{j=1}^{m} \beta\left(x_{ij}|z(k, y_{ij}), k - z(k, y_{ij}) + 1\right) \tag{12}
$$

where $\rho_1 = v_1$ and $\rho_s = v_s \prod_{l=1}^{s-1}(1 - v_l)$ for $l = 2, 3, \ldots$ where $v_s \sim \beta(1, M)$ and $\mathbf{y}_i$ follows

the baseline distribution F_0 and

$$z(k,\cdot) = \sum_{r=1}^{k} r\delta_{(r-1)/k,r/k]}(\cdot)$$

and δ is the Dirac delta function.

Following Walker (2007) , we introduce a uniform latent variable to convert the infinite mixture representation into a finite mixture representation as follows:

$$\begin{aligned} f(\mathbf{x}_i, u_i|k,\boldsymbol{\rho},\mathbf{y}) &= \sum_{s=1}^{\infty} \mathbf{1}(u_i < \rho_s) \prod_{j=1}^{m} \beta\left(x_{ij}|z(k,y_{sj}), k - z(k,y_{sj}) + 1\right) \\ &= \sum_{s\in A_{\boldsymbol{\rho}}(u_i)} \prod_{j=1}^{m} \beta\left(x_{ij}|z(k,y_{sj}), k - z(k,y_{sj}) + 1\right) \end{aligned}$$

where the set $A_{\boldsymbol{\rho}}(u_i) = \{s : u_i < \rho_s\}$ which is clearly a finite set.

Finally, we introduce a further latent label variable as follows:

$$f(\mathbf{x}_i, u_i, d_i|k,\boldsymbol{\rho},\mathbf{y}) = \mathbf{1}(u_i < \rho_{d_i}) \prod_{j=1}^{m} \beta\left(x_{ij}|z(k,y_{d_ij}), k - z(k,y_{d_ij}) + 1\right)$$

Therefore, the complete likelihood function based on the sample $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is:

$$l(k,\boldsymbol{\rho},\mathbf{y}|\mathbf{X},\mathbf{u},\mathbf{d}) = \prod_{i=1}^{n}\prod_{j=1}^{m} \mathbf{1}(u_i < \rho_{d_i})\beta\left(x_{ij}|z(k,y_{d_ij}), k - z(k,y_{d_ij}) + 1\right).$$

where $\mathbf{u} = (u_1, u_2, \ldots, u_n)$ and $\mathbf{d} = (d_1, d_2, \ldots, d_n)$.

Then an MCMC algorithm to sample the posterior parameter distribution can be set up as follows:

1. Set an initial allocation $\mathbf{d} = (d_1, \ldots, d_n)$
2. Generate v_s by simulating from $v_s \sim \beta(n_s+1, n-\sum_{l=1}^{s} n_l+M)$ for $s = 1, \ldots, d^*$, where $d^* = \max\{d_i : i = 1, 2, \ldots, n\}$ and $n_s = \sum_{i=1}^{n} \mathbf{1}(d_i = s)$.
3. Update u_i by simulating from $U(1, \rho_{d_i})$ for $i = 1, \ldots, n$.

4. Update v_s by simulating from $v_s \sim \beta(1, M)$ for $s = d^* + 1, \ldots, s^*$, where s^* is such that $\sum_{s=1}^{s^*} \rho_s > 1 - u^*$.

5. Update $\{y_{sj}\}_{j=1}^m$ by simulating the following full conditional distribution for $s = 1, \ldots, s^*$.

$$f(y_{sj}|\cdots) \propto \prod_{i:d_i=s} \beta(x_{ij}; z(k, y_{sj}), k - z(k, y_{sj}) + 1). \tag{13}$$

If there is no d_i equal to s then $y_{sj}|\cdots \sim F_0$.

6. Update d_i, $i = 1, \ldots, n$, by simulating from

$$P(d_i = s|\cdots) \propto \mathbf{1}(u_i < \rho_s) \prod_{j=1}^{m} \beta(x_{ij}; z(k, y_{sj}), k - z(k, y_{sj}) + 1) \tag{14}$$

7. Update k by simulating the following full conditional distribution

$$P(k|\cdots) \propto p(k) \prod_{i=1}^{n} \prod_{j=1}^{m} \beta(x_{ij}; z(k, y_{d_i j}), k - z(k, y_{d_i j}) + 1) \tag{15}$$

5 Simulations and Empirical Applications

In this section, we undertake several simulation studies and a real data example to illustrate the performance of the proposed nonparametric Bayesian approach. For simplicity in the visualization, we only consider examples in the two-dimensional case in order to better illustrate the accuracy in density estimation.

5.1 Simulated data

We consider simulated data from the bivariate beta distribution proposed in Olkin and Liu (2003). This is a continuous variable with support on the unit square and it is a generalization of the univariate beta distribution function to the bivariate case. The bivariate beta distribution, $\beta_2(x, y; a, b, c)$, is derived by considering the joint distribution of two random

variables:

$$X = \frac{U}{U+V}, \quad Y = \frac{V}{V+W},$$

where U, V and W are three independent standard gamma distributions with respective shape parameters, a, b and c. Clearly, the marginal distributions of X and Y are beta distributions, $\beta(x; a, c)$ and $\beta(y; b, c)$, respectively. This model can describe a wide rage of densities on the unit square and can be easily generalized to the multivariate case.

As a first example, we consider 200 data simulated from a bivariate beta distribution $\beta_2(x, y; 5, 10, 10)$. For these data, we apply the proposed bayesian density estimation method based on Bernstein polynomials. We impose the following noninformative prior assumptions. For the baseline distribution, F_0, we assume a uniform distribution on $[0, 1]^2$. We also set the smoothing parameter to be $M = 1$, as suggested in Petrone (1999b), in order to express a small degree of belief in the prior guess. Finally, we assume a uniform prior distribution for k in the interval [0,100]. The proposed MCMC algorithm described in Section 4 is run using 10000 iterations and discarding the first 1000 as burn-in iterations. Figure 1 shows the estimated predictive and true density for these data. The predictive and true marginal densities are also shown. We can see that the Bernstein polynomial density model provides a good fit to the data.

In order to illustrate the flexibility of the model, we now consider 200 simulated data from a mixture of bivariate beta densities $\beta_2(x, y; 5, 10, 10)$ and $\beta_2(x, y; 5, 1, 5)$ with equal weights. Using the same prior specifications as before, the proposed MCMC algorithm is run for these data using the same number of iterations. The true and predictive joint densities are illustrated in Figure 2. The marginal predictive and true densities corresponding to mixtures of univariate beta distributions are also shown. We can observe that also in this mixture case there is is a good fit to the true densities.

We have also tried alternative prior specifications. In general, there is little sensitivity of the density estimations to the choice of the concentration parameter M. Similar to the univariate case in Petrone (1999b), the predictive densities get somewhat closer to the uniform prior distribution for larger values of M. On the other hand, as we would expect, there is slightly more sensitivity to the prior specification for k. We have observed that using

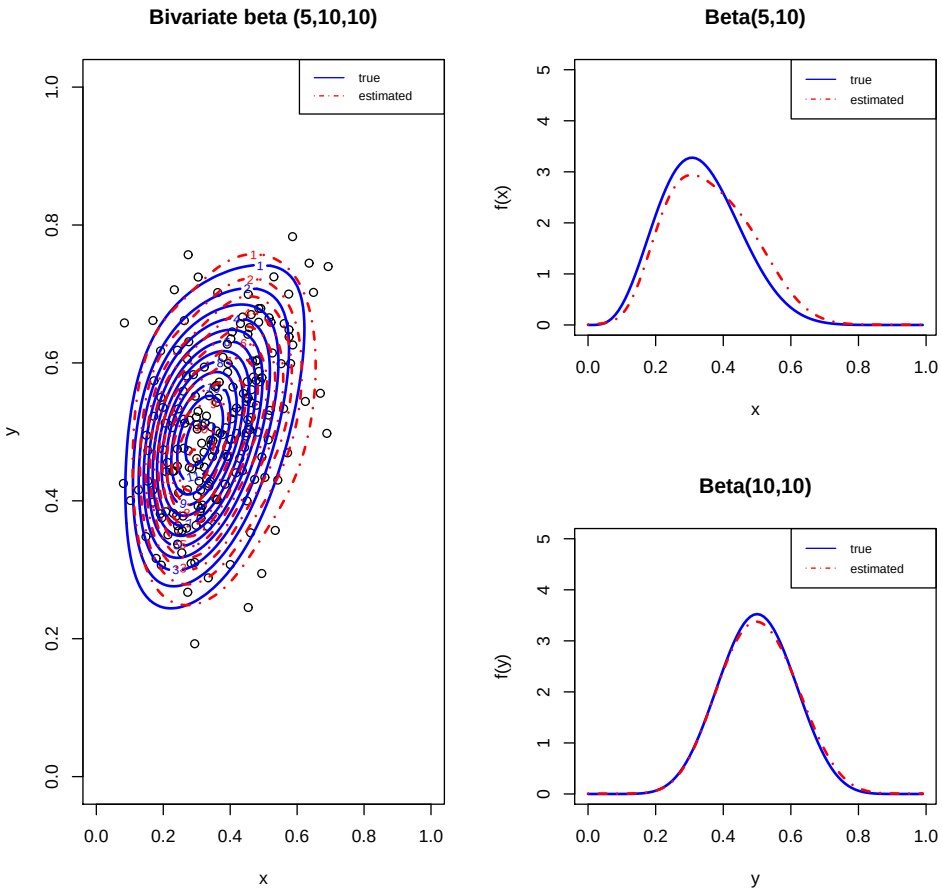

Figure 1: Predictive and true densities obtained from 200 simulated data from a bivariate beta distribution $\beta_2(x, y; 5, 10, 10)$ (right) and marginal distributions (left).

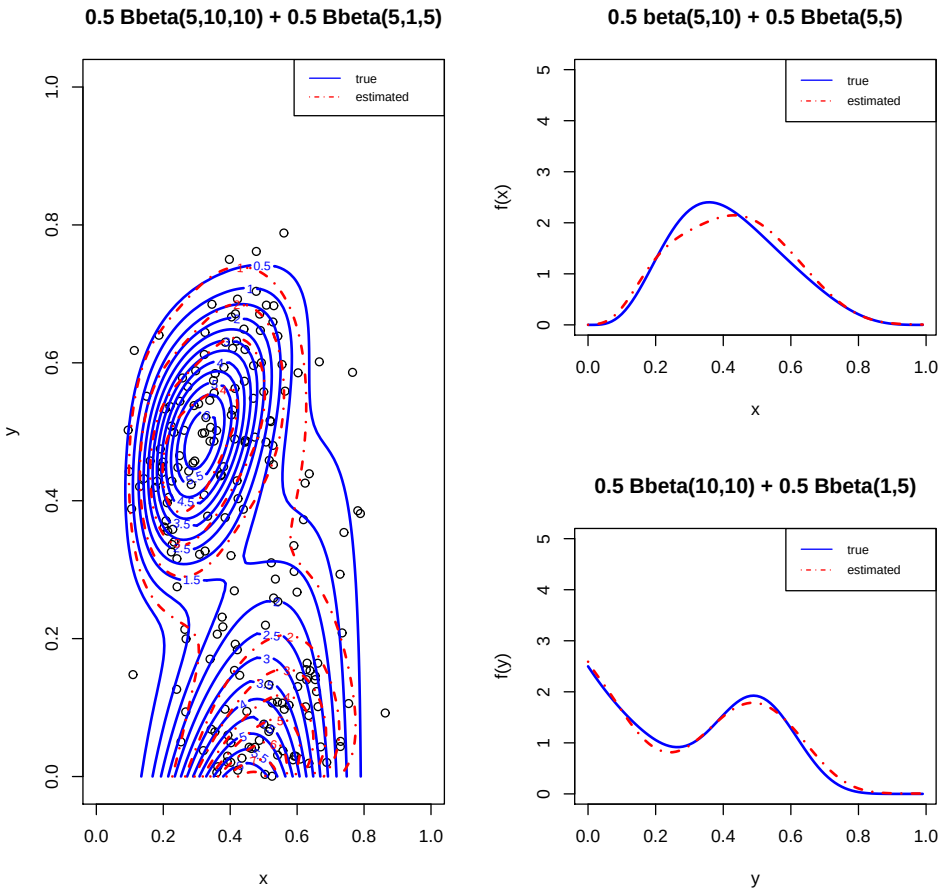

Figure 2: Predictive and true densities obtained from 200 simulated data from a mixture of bivariate beta distributions $\beta_2(x, y; 5, 10, 10)$ and $\beta_2(x, y; 5, 1, 5)$ with equal weights (right) and marginal distributions (left).

prior distributions for k concentrated on small values, such as a Poisson prior distribution with small mean, lead to smoother predictive densities than using a uniform prior on a closed interval. This is also observed for the univariate Bernstein model in Petrone (1999b). As noted earlier, k plays a similar role in the Bernstein polynomial to the bandwidth in kernel density estimation.

Finally, it is important to note that the computational cost of the method is not high and its application is feasible in practice. In these two examples, the computing time was around half an hour for each case using self programmed code in R 2.15.2 (R Development Core Team 2011) on a computer with a 3.4 Ghz core. In contrast, we also programmed a Polya urn type sampling scheme as in Petrone (1999b) which took over an hour for these two examples.

5.2 Real data example

In this section, we examine the relationship between the percentage of forest area (% of land area) and percentage of agricultural nitrous oxide emissions (% of total) in 127 countries in 2010. The data are available from `http://data.worldbank.org/`. Nitrous oxide is naturally present in the atmosphere, however, human activities in agriculture such as fertilizer use and waste and savannah burning are increasing the amount of this gas in the atmosphere. The impact of nitrous oxide emissions on warming the atmosphere is over 300 times that of carbon dioxide per unit weight. Therefore, it is interesting to examine the influence of the percentage forest area in the reduction of these emissions.

Figure 3 shows the scatter plot of these data together with the estimated joint density using the proposed Bernstein polynomial model with the same prior assumptions and MCMC iterations as in the simulation examples. We can observe that the model identifies three main clusters of countries. Firstly, there is a large group corresponding to those countries with more than 10% of forest area where there is a clear negative relationship between the percentage of forest area and the nitrous oxide emissions. Secondly, there is a fairly large group with less than 10% of forest area but comparatively large percentage of nitrous oxide emissions. Finally, there is a small group of countries with low percentage of forest area

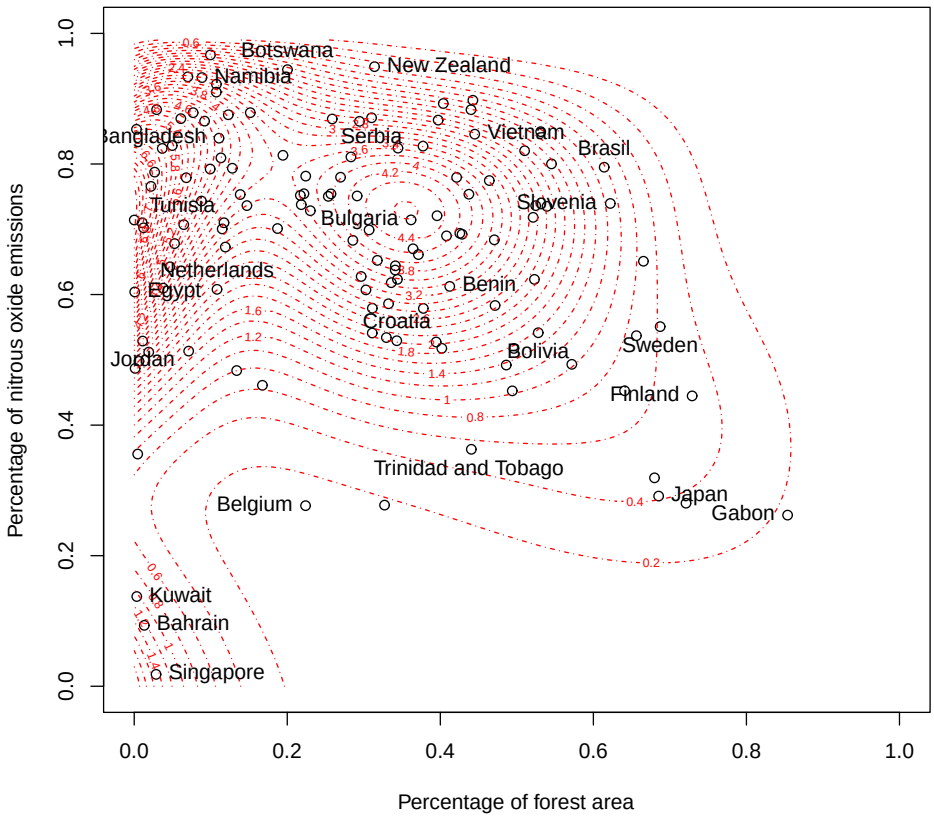

Figure 3: Predictive joint density of the percentage of forest area and percentage of agricultural nitrous oxide emissions obtained from a data base of 127 countries in 2010.

and low percentage of nitrous oxide emissions d.

Finally, Figure 4 shows the estimated marginal distributions of the percentage of forest area and the percentage of nitrous oxide emissions. We can observe that the distribution of the percentage of forest area has two modes, one is zero and the other is close to 0.4. The nitrous emissions percentage distribution is left-skewed with a mode close to 0.8. It seems that the model is flexible enough to capture adequately the different shapes of these distributions.

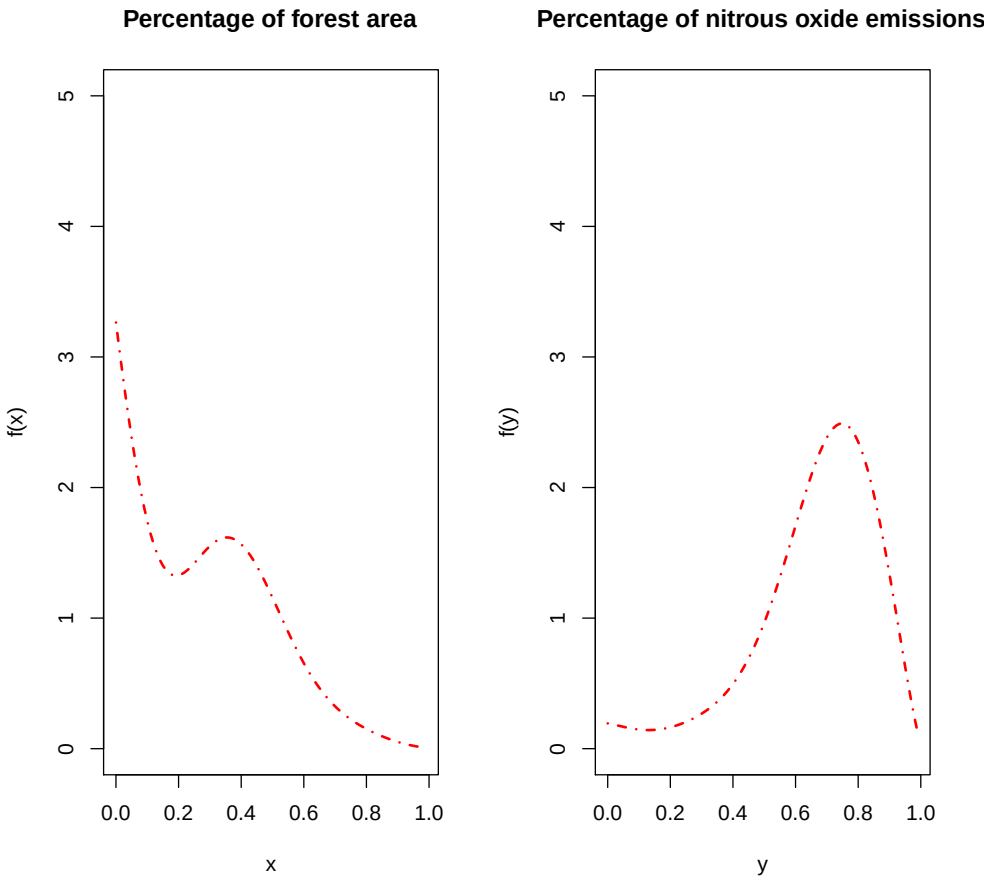

Figure 4: Marginal estimated distribution of the percentage of forest area (left) and percentage of agricultural nitrous oxide emissions (right) obtained from a data base of 127 countries in 2010.

6 Conclusions and extensions

In this paper, we have extended the Bernstein Dirichlet prior introduced in Petrone (1999a,b) for densities on a closed interval to the multivariate case and have obtained the convergence rate of the associated posterior distribution. Moreover, we have introduced a new algorithm for sampling from the posterior distribution. Various extensions are possible.

Firstly, although here we have defined the multivariate Bernstein polynomial using a single k, in principle it is possible to consider different values $k_1, \ldots, k_m$ for the different components of $\mathbf{x}$. This might be useful from a practical viewpoint if some variables are more spread than others. Secondly, following Tenbusch (1994), it would be interesting to consider multivariate Bernstein densities on the triangle which might be more appropriate for modeling the joint density of various proportions of the same quantity. Finally, the multivariate Bernstein polynomial provides an asymptotic model for a copula, see e.g. Sancetta and Satchell (2004) so that it can be used to model the dependence structure of a multivariate distribution. Then the use of a Bernstein Dirichlet prior for the copula could be combined with standard, Bayesian nonparametric priors for the marginals to provide a general, nonparametric approach to multivariate data modeling.

References

Ghosal, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *Ann. Stat. 29*, 1264–1280.

Ghosal, S., J. Ghosh, and A. Van Der Vaart (2000). Convergence rates of posterior distributions. *Annals of Statistics 28*(2), 500–531.

Hjort, N., C. Holmes, P. Müller, and S. Walker (2010). *Bayesian Nonparametrics*. Cambridge: Cambridge University Press.

Jara, A., T. Hanson, F. Quintana, P. Müller, and G. Rosner (2011). Dppackage: Bayesian semi- and nonparametric modeling in r. *J. Stat. Softw. 40*, 1–30.

Lorentz, G. (1986). *Bernstein Polynomials*. New York: Chelsea.

Olkin, I. and R. Liu (2003). A bivariate beta distribution. *Statistics & Probability Letters 62*(4), 407–412.

Papaspiliopoulos, O. and G. Roberts (2008). Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika 95*(1), 169–186.

Petrone, S. (1999a). Bayesian density estimation using Bernstein polynomials. *Can. J. Stat. 27*, 105–126.

Petrone, S. (1999b). Random Bernstein polynomials. *Scand. J. Stat. 26*, 373–393.

Petrone, S. and L. Wasserman (2002). Consistency of Bernstein polynomial posteriors. *J. Roy. Stat. Soc. B 64*, 79–100.

Phillips, G. (2003). *Interpolation and Approximation by Polynomials*. New York: Springer.

Sancetta, A. and S. Satchell (2004). The Bernstein copula and its applications to modeling and approximations of multivariate distributions. *Economet. Theor. 20*, 535–562.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*, Volume 26. London: Chapman & Hall.

Tenbusch, A. (1994). Two-dimensional Bernstein polynomial density estimators. *Metrika 41*, 233–253.

Trippa, L., P. Bulla, and S. Petrone (2011). Extended bernstein prior via reinforced urn processes. *Ann. Inst. Statist. Mat. 63*, 481–496.

Vitale, R. (1975). A Bernstein polynomial approach to density function estimation. *Statistical Inference and Related Topics 2*, 87–99.

Walker, S. (2007). Sampling the dirichlet mixture model with slices. *Communications in Statistics: Simulation and Computation® 36*(1), 45–54.

Appendix

Proof of Theorem 3.3

For $k \geq 1$, define $f_k(x_1, x_2, \ldots, x_m) = b(x_1, x_2, \ldots, x_m; k, F_0)$, where F_0 is the cumulative distribution function for f_0. Note that f_k is also uniformly bounded away from 0 for all large k by (5). Note that we may also write $f_k(x_1, x_2, \ldots, x_m) = b(x_1, x_2, \ldots, x_m; k, \mathbf{w}_k^0)$, where $\mathbf{w}_k^0 = \{w_{i_1 i_2 \ldots i_m, k}\}_{i_l=0,\ l=1,2,\ldots,m}^{k}$ and for $i_l = 1, 2, \ldots, k, l = 1, 2, \ldots, m$

$$w_{i_1 i_2 \ldots i_m, k} = \int_{(i_1-1)/k}^{i_1/k} \cdots \int_{(i_m-1)/k}^{i_m/k} f_0(x_1, x_2, \ldots, x_m)\, \mathrm{d}x_1\, \mathrm{d}x_2 \ldots\, \mathrm{d}x_m$$

Also observe that

$$\begin{aligned}
&\mid b(x_1, x_2, \ldots, x_m; k, \mathbf{w}_k) - b(x_1, x_2, \ldots, x_m; k, \mathbf{w}_k^0) \mid \\
&= \left| k^m \sum_{i_1=1}^{k} \cdots \sum_{i_m=1}^{k} (w_{i_1 i_2 \ldots i_m, k} - w^0_{i_1 i_2 \ldots i_m, k}) \binom{k-1}{i_1 - 1} \cdots \binom{k-1}{i_m - 1} x_1^{i_1} (1-x_1)^{k-i_1} \ldots x_m^{i_m} (1-x_m)^{k-i_m} \right| \\
&\leq k^m \max_{1 \leq i_1, \ldots, i_m \leq k} |w_{i_1 i_2 \ldots i_m, k} - w^0_{i_1 i_2 \ldots i_m, k}| \\
&\leq k^m \sum_{i_1=1}^{k} \cdots \sum_{i_m=1}^{k} \left| w_{i_1 i_2 \ldots i_m, k} - w^0_{i_1 i_2 \ldots i_m, k} \right| \\
&= k^m \|\mathbf{w}_k - \mathbf{w}_k^0\|_1
\end{aligned}$$

Therefore if $\|\mathbf{w}_k - \mathbf{w}_k^0\|_1 \leq \varepsilon^{m+1}$ and $d_1 \varepsilon^{-1} \leq k \leq d_2 \varepsilon^{-1}$ for some constants d_1 and d_2, then $\sup_{0 < x_1, x_2, \ldots, x_m \leq 1} |f(x_1, x_2, \ldots, x_m) - b(x_1, x_2, \ldots, x_m; k, \mathbf{w}_k)| \leq D_1 \varepsilon$ for a constant D_1 and also $b(x_1, \ldots, x_m; k, \mathbf{w}_k)$ is bounded away from 0 for sufficiently small ε. It therefore follows that for some constant D_2, $h(f_0, b(\cdot, \cdots, \cdot; k, \mathbf{w}_k)) \leq D_2 \varepsilon$ and so (8.6) of Ghosal et al. (2000) implies that $b(\cdot, \cdots, \cdot; k, \mathbf{w}_k) \in N(C_1 \varepsilon, f_0)$ for a constant C_1. Hence

$$N(C_1 \varepsilon, f_0) \supset \{b(x_1, \ldots, x_m; k, \mathbf{w}_k) : \|\mathbf{w}_k - \mathbf{w}_k^0\|_1 \leq \varepsilon^{m+1}\}.$$

If we choose k_n satisfying

$$b_1 \left(\frac{n}{\log n}\right)^{1/(m+2)} \leq k_n \leq b_2 \left(\frac{n}{\log n}\right)^{1/(m+2)}$$

for some constants b_1 and b_2 and $\tilde{\varepsilon}_n = k_n^{-1}$, Lemma A.1 of the Appendix in Ghosal (2001) implies that for some constants C_3, C_4, D and d,

$$\begin{aligned}
\Pi(N(C_1\tilde{\varepsilon}_n, f_0)) &\geq \rho(k_n) C_2 \exp(-C_3 k_n^m \log(1/\tilde{\varepsilon}_n)) \\
&\geq B_1 \exp(-\beta_1/\tilde{\varepsilon}_n) \times C_2 \exp(-C_3 (1/\tilde{\varepsilon}_n)^m \log(1/\tilde{\varepsilon}_n)) \\
&\geq D \exp(-d(1/\tilde{\varepsilon}_n)^m \log(1/\tilde{\varepsilon}_n))
\end{aligned}$$

Hence $\tilde{\varepsilon}_n = n^{-1/(m+2)} (\log n)^{1/(m+2)}$ satisfies condition (9) of Theorem 3.1.

Let s_n be an integer satisfying

$$L_1 (1/\tilde{\varepsilon}_n)^m \log(1/\tilde{\varepsilon}_n) \leq s_n \leq L_2 (1/\tilde{\varepsilon}_n)^m \log(1/\tilde{\varepsilon}_n)$$

for some constants L_1 and L_2. Then

$$L_1^{'}\ n^{m/(m+2)} (\log n)^{2/(m+2)} \leq s_n \leq L_2^{'}\ n^{m/(m+2)} (\log n)^{2/(m+2)}$$

where we may choose $L_1^{'} = \frac{L_1}{2m+4}$ and $L_2^{'} = \frac{L_2}{m+2}$. Put $\mathscr{F}_n = \bigcup_{r=1}^{s_n} \mathscr{B}_r$. Then for constants B_3, B and L,

$$\Pi(\mathscr{F}_n^c) \leq \sum_{r=s_n+1}^{\infty} \rho(r) \leq B_3 \exp(-\beta_2 s_n) \leq B \exp(-L(1/\tilde{\varepsilon}_n)^m \log(1/\tilde{\varepsilon}_n))$$

and L can be made arbitrarily large by choosing L_1 sufficiently large. As $(1/\tilde{\varepsilon}_n)^m \log(1/\tilde{\varepsilon}_n)$ and $n\tilde{\varepsilon}_n^2$ have the same order, the condition (8) holds. Now by the arguments given in the

proof of Theorem 3.2, for some constants C and L_3, we have

$$\begin{aligned}\log D(\varepsilon, \mathscr{F}_n, d) &\leq s_n \log(C/\varepsilon) + \log(s_n) \\ &\leq L_2^{'}\ n^{m/(m+2)} (\log n)^{2/(m+2)} \log(C/\varepsilon) + \log(L_2^{'}\ n^{m/(m+2)} (\log n)^{2/(m+2)}) \\ &\leq L_3\ n^{m/(m+2)} (\log n)^{2/(m+2)} \log(C/\varepsilon)\end{aligned}$$

So (7) holds for the choice $\bar{\varepsilon}_n = n^{-1/(m+2)} (\log n)^{(m+4)/(2m+4)}$. Hence the posterior converges at the rate $n^{-1/(m+2)} (\log n)^{(m+4)/(2m+4)}$.